Abstract Title Page

Title: Developing a theory of treatment effect heterogeneity through better design: Where do behavioral science interventions work best?

Authors and Affiliations:
Elizabeth Tipton, *Teachers College, Columbia University*
David Yeager, *University of Texas – Austin*
Ronaldo Iachan, *ICF*

Abstract Body

Background / Context:

Recently, President Obama issued an executive order calling policy makers throughout the government to use research findings from behavioral science to design policies that better serve the American people (Executive Order, September 15, 2015). This call refers to recent findings in behavioral economics and psychology focused on the promise of brief, scalable, low-cost interventions (e.g., Thaler & Sunstein, 2008; Walton, 2014; Wilson & Juarez, 2015; Yeager & Walton, 2011). In education, these behavioral science interventions have included both 'nudges'– for example, text-message based interventions (e.g., Castleman and Page, 2014; York and Loeb; 2014) – as well as interventions aimed at changing 'mindsets' – like learning that the brain is a muscle that can grow and develop (e.g., Paunesku et al, 2015). While the results from these studies are promising – thus leading to the executive order – to date little is known about how well the results of these interventions may generalize, or, as a corollary, the conditions under which they may work best. Knowing this is essential for the ethical and responsible use of behavioral insights in policy.

Questions regarding the generalizability of results from educational experiments have been at the forefront of methods development over the past five years. This work has focused on methods for estimating the effect of an intervention in a well-defined inference population (e.g., Tipton, 2013; O'Muircheartaigh and Hedges, 2014); methods for assessing similarity between the students and schools in a study to those in an inference population (e.g., Stuart, Cole, Bradshaw, and Leaf, 2011; Tipton, 2014a); and methods for improved site selection with generalization in mind (e.g., Tipton et al, 2014; Tipton, 2014b). To date, this work has been developed in the context of the large-scale education experiments typically funded by IES, wherein schools are typically randomly assigned to receive a yearlong curricular intervention. Given this context, this work on generalization has focused largely on estimation of the *average treatment effect* in a population, particularly under the assumption that random selection of schools into the study is infeasible (see Olsen, Orr, Bell, and Stuart, 2013).

In this paper, we argue that behavioral science interventions are different in two important ways from these standard large-scale education experiments, and that these differences provide an opportunity to develop new methods for generalization. First, behavioral science interventions are more often randomly assigned to students, not to intact schools; this results in the more powerful multi-site (i.e., random-block) design that in addition to an average treatment impact, also allows for estimation of the distribution of treatment impacts across schools. Second, and perhaps of even greater importance, the fact that these behavioral science interventions do not focus on curricular changes and do not require large and lasting changes to school routines means that it is easier to recruit schools and students into the studies. In the paper we argue that these differences allow researchers to shift from answering only questions about the average effect, to questions regarding variability in treatment impacts as well.

Purpose / Focus of Study:

This paper focuses on design considerations for those developing behavioral science intervention experiments with a focus on the development of methods for making generalizations from these studies. We begin by arguing that the brevity and affordability of the interventions makes it possible for a dual-randomization procedure to be implemented; in this design, schools are randomly selected into the study from a well-defined population frame and then students

within these schools are randomly assigned to the intervention. This dual-randomization procedure enables a model-free estimation of the *population* average treatment effect. While statistically ideal, this procedure has been rarely enacted in randomized experiments (see Olsen et al., 2013), and in this paper, we argue that it should be more commonly attempted in behavioral science intervention studies.

Second, we argue that the fact that the typical behavioral science intervention study involves random assignment to students within schools also means that questions of treatment impact heterogeneity become not only important, but also estimable. If treatment impacts vary – which is the assumption motivating the work on generalization issues to date – then the *average* treatment impact is not sufficient. Instead, we argue that the ideal behavioral science intervention study should also aim to answer questions regarding the development of a *theory of treatment impact heterogeneity*. This amounts to developing a design in which two additional generalization questions can be answered: To what extent does the treatment impact vary across schools in the population? Does the treatment impact vary in relation to pre-specified school or student features?

The paper focuses on the development of a study design aimed at answering all three of these generalization questions – regarding the average effect, its variability, and treatment effect moderation. The concern here has to do with the development of the *optimal* design for these studies, wherein all three estimands are adequately powered. Importantly, this involves trade-offs, since, as we will show, the design that is most powerful for estimating the average treatment effect also results in less power for estimation of moderator effects (and vice versa).

The NGMI Experiment/ Setting/ Population/ Intervention:

Throughout the paper we develop the statistical theory regarding generalization in relation to the process of creating a probability-sampling plan for the National Growth Mindset Intervention [NGMI] study. The NGMI study evaluates the effect of a brief psychological intervention based on research on *growth mindset* developed by Dweck (2006), Yeager and Dweck (2012) and others (Paunesku et al., 2015). In this section, we briefly provide background information on the study, since it offers motivation for the generalization questions herein.

Setting. The study takes place in U.S. public 'regular' high schools. This includes only schools serving grades 9 – 12, and excludes: charter and magnet schools; schools serving special populations (e.g., Bureau of Indian Affairs schools); and schools with fewer than 25 9th grade students.

Population. The target population is all 9th grade students attending the 9,190 'regular' U.S. public high schools (as defined above). The sample is a national probability sample including all 9th grade students from 80 of these schools.

Intervention. The study design involves two 25-minute sessions of an on-line growth mindset intervention designed to communicate the message that intelligence is malleable and that individual intellectual abilities can be increased through deliberate effort and practice (see Figure 1). The interventions draws on decades of research on attitudes and behavior change in order to communicate these ideas effectively in a short span of time by having students read information, answer questions, and complete writing exercises.

Outcomes. Survey data are collected at each of the two online sessions. Students are followed into the beginning of the first semester of their 10th grade year (and possibly thereafter) with administrative records providing outcome information, including grades, test scores (when available), and measures of completion (e.g., course taking in math).

Significance / Novelty of study:

The question at the heart of the NGMI study – and we argue, all behavioral science intervention studies – is how to design the study to not only estimate the average treatment effect, but also variability in impacts, and to develop a theory of treatment effect heterogeneity. The approach we develop in the paper fuses together results from the survey sampling and experimental design literatures. It requires researchers to develop hypotheses about treatment effect heterogeneity at the outset of the study, to develop methods to measure these moderators, and to use a stratified selection procedure, with strata based on potential moderators.

Importantly, while stratified selection is common in survey sampling, the approach developed here deviates from standard practice in that the focus is on strata based on covariates explaining variation in treatment impacts, *not* outcomes. For example, it is common in probability sampling to stratify schools based on a measure such as free-and-reduced-lunch status – a variable often highly correlated with outcomes; we show here that this approach makes little sense, unless the intervention effects also vary in relation to this variable. Similarly, the approach developed here differs from standard results in experimental design with regard to power. For example, standard results by Raudenbush and Liu (2000) focus on the power for each of the three parameters (i.e., average, variation, moderation) under study *separately*, whereas in our approach power must be considered for what we call 'multipurpose' designs.

Statistical, Measurement, or Econometric Model

Stratification

In the 'multipurpose' design developed in this paper, researchers must begin their study by identifying possible moderators; these should be based upon previous research, the theory of change, and on questions about context that may be of particular interest to policy makers. This list of moderators could easily be large, in which case dimension reduction methods may be required[1]. For example, this could include the use of *k*-means cluster analysis (see Tipton, 2014b), wherein moderators that co-vary are stratified together; an alternative method (which is used in the NGMI study) is to use a latent-variable approach.

An important consideration in designing strata is the concern that certain features of the population may co-vary to an extent that, without prior planning, the effects are essentially *aliased* together. For example – as found in the NGMI study – school achievement and school minority composition can be highly correlated ($r > .6$). In these cases, strata can be used to de-alias the effects; this is particularly important since given their high correlation, without planning, it may be difficult to impossible to separate these effects well using post-hoc methods alone. Notably, concerns of this type are addressed commonly in the experimental design literature, but are rarely found in the survey sampling; this speaks to the need for these new 'multipurpose' designs.

In the remainder of this section, assume that we have classified the population in $k = 1 \ldots M$ strata, $S = \{S_1, S_2, \ldots, S_M\}$, and that each stratum S_j contains N_j units in the population. In Table 1 (see Appendix) we illustrate this for the NGMI study; here $M = 5$. In this framework, the 'multipurpose' design question becomes: how should the total sample n be best allocated to these M strata (i.e., how do we determine n_j)?

[1] This is because stratifying on p covariates leads to at least 2^p strata, which can be far too many given the small number of clusters.

HLM formulation

One strategy for allocating the sample to the strata is to use proportional allocation (i.e., $n_j = n^*N_j/N$), as illustrated in the second row of Table 1. In this case, when estimating the average treatment effect and variation in the population, a standard random-block design HLM model (see Raudenbush & Liu, 2000) could be used. In this paper, we show that this 'proportional-allocation' design can be far from optimal for estimating moderator relationships, particularly if any of the strata are small. In Table 1, for example, power for estimating differences in treatment impacts between Strata 4 and 5 would be compromised, given the very small proportion of the population in Stratum 5.

Given the multiple estimands, we argue that typically non-proportional allocation will be desirable in order to achieve enough power for estimating both moderator relationships and average treatment impacts and variation. An example of this is given in the third row of Table 1; note here that now the proportion in Stratum 5 has been increased, while the proportion in other strata has been reduced.

When using a non-proportional allocation scheme, stratified estimators of the average treatment effect and treatment effect variation are required. To do so, first a stratified HLM model must be estimated,

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2)$$
$$\beta_{0jk} = \sum\gamma_{0k}S_{jk} + u_{0jk} \qquad u_{0jk} \sim N(0, \tau_0^2)$$
$$\beta_{1jk} = \sum\gamma_{1k}S_{jk} + u_{1jk} \qquad u_{1jk} \sim N(0, \tau_1^2)$$

where here $\boldsymbol{\gamma}_1 = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1M})$ are the stratum average treatment impacts for strata $k = 1 \ldots M$. Note here that τ_w^2 is the variation in treatment impacts *within* strata. This means that now the population average treatment impact is defined as a weighted combination of these,

$$\mu_s = \Sigma w_k\gamma_{1k}$$

where $w_k = N_k/N$. The population variation in treatment impacts (τ_t^2) is defined as,

$$\tau_t^2 = \tau_w^2 + \tau_b^2,$$

where τ_w^2 is the pooled within-stratum variation in treatment impacts, and τ_b^2 is the between-stratum variation in treatment impacts, which we can define as

$$\tau_b^2 = \Sigma(\gamma_{1k} - \mu_s)^2/(M - 1).$$

In the paper, we provide results regarding estimators of these parameters, and the effect of post-stratification on power.

Usefulness / Applicability of Method:

Throughout the paper, we situate the discussion of the development of the 'multipurpose' design in relation to the NGMI study. In this study, for example, the strata are created in relation to two variables: 1) school achievement, which we create using a latent variables approach based on data from GreatSchools.org and the College Board; and 2) school minority composition. This resulted in a design with 5 strata, which is indicated in Table 1. The final allocation used for sampling is found in the third row of Table 1.

Conclusions:

Behavioral science interventions are becoming more common in education and policy more generally. With this comes the opportunity to develop generalizable theories of treatment effect heterogeneity—something that is difficult to do with whole-school reform. By doing so, researchers will be better situated to help policy makers and school officials understand where these brief interventions hold most promise, and where further research is needed.

Appendices

Appendix A. References

Castleman, B. L., & Page, L. C. (2014). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*.

Dweck, C. (2006). *Mindset: The new psychology of success*. Random House, Chicago, IL.

Executive Order, September 15, 2015. Retrieved on October 1, 2015 from https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american.

Olsen, R.B., Orr, L.L., Bell, S.H., and Stuart, E.A. (2013) External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32* 107-121.

O'Muircheartaigh, C., Hedges L.V. (2014) Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society, Series C* (Applied Statistics), 63: 195-210.

Paunesku, D., Walton, G.M., Romero, C.L., Smith, E.N., Yeager, D.S., & Dweck, C.S. (2015). Mindset interventions are a scalable treatment for academic underachievement. *Psychological Science, 26,* 784-793.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological methods*, *5*(2): 199 – 221.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, Part 2, 369-386.

Thaler, R.H. & Sunstein, C.R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, New Haven.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*: 239-266.

Tipton, E. (2014a) How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics, 39*(6): 478 – 501.

Tipton, E. (2014b) Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review, 37*(2): 109-139.

Tipton, E., Hedges, L.V., Vaden-Kiernan, M., Borman, G.D., Sullivan, K. & Caverly, S. (2014) Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness, 7*(1): 114-135.

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*(1), 73-82.

Wilson, T.D., & Juarez, L.P. (2015) Intuition is not evidence: Prescriptions for behavioral interventions. *Behavioral Science and Policy, 1(1)*.

Yeager, D.S. & Dweck, C.S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47,* 302-314.

Yeager, D.S. & Walton, G. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81,* 267-301.

York, B.N. & Loeb, S. (2014) One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents of Preschoolers. National Bureau of Economic Research Working Paper. Working Paper Series 20659.

Appendix B. Tables and Figures

Figure 1: Student administration of the intervention

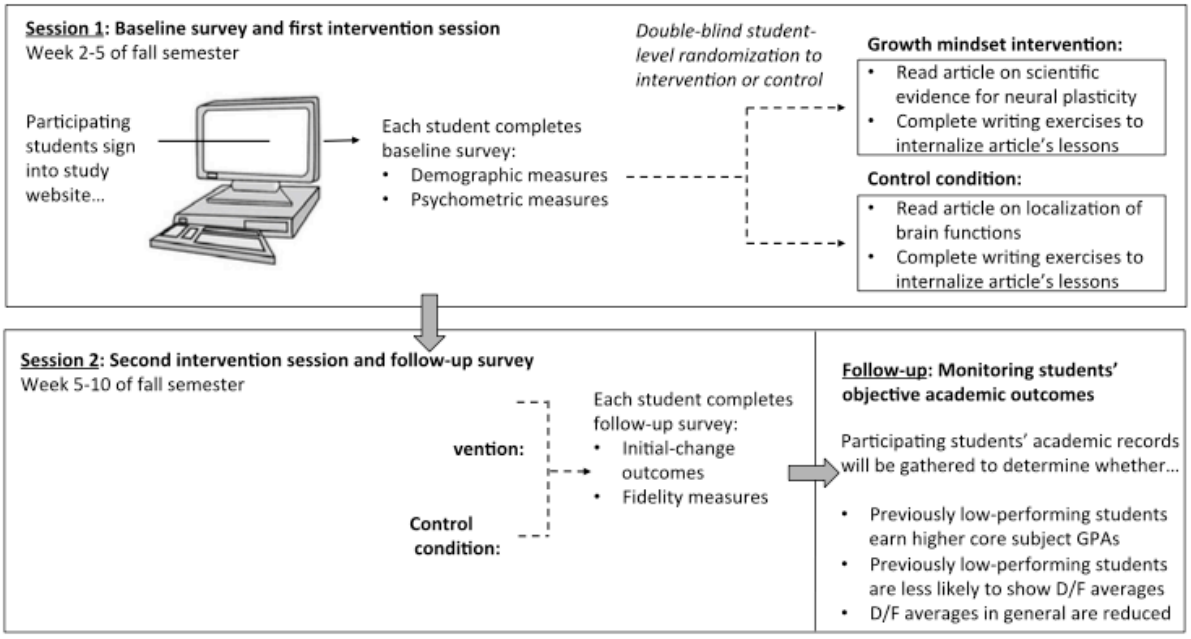

Table 1: Strata and allocation schemes in the NGMI study

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Population (out of N = 9,900)	25%	27%	23%	20%	5%
Proportional Allocation - (out of *n* = 100)	25	27	23	20	5
Multipurpose Allocation - (out of *n* = 100)	20	24	23	14	19